

Automatic Inference of Graph Transformation Rules Using the Cyclic Nature of Chemical Reactions

Christoph Flamm^{2,8}, Daniel Merkle (✉)¹, Peter F. Stadler^{2,7}, and
Uffe Thorsen¹

¹ Department of Mathematics and Computer Science, University of
Southern Denmark, Odense M DK-5230, Denmark, {daniel,
uthorsen}@imada.sdu.dk

² Institute for Theoretical Chemistry, University of Vienna, Wien
A-1090, Austria, xtof@tbi.univie.ac.at

³ Bioinformatics Group, Department of Computer Science, and
Interdisciplinary Center for Bioinformatics, University of Leipzig,
Leipzig D-04107, Germany, stadler@bioinf.uni-leipzig.de

⁴ Max Planck Institute for Mathematics in the Sciences, Leipzig
D-04103, Germany

⁵ Fraunhofer Institute for Cell Therapy and Immunology, Leipzig
D-04103, Germany

⁶ Center for non-coding RNA in Technology and Health,
University of Copenhagen, Frederiksberg C DK-1870, Denmark

⁷ Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe NM 87501,
USA

⁸ Research Network Chemistry Meets Microbiology, University of
Vienna, Wien A-1090, Austria

Abstract

Graph transformation systems have the potential to be realistic models of chemistry, provided a comprehensive collection of reaction rules can be extracted from the body of chemical knowledge. A first key step for rule learning is the computation of atom-atom mappings, i.e., the atom-wise correspondence between products and educts of all published chemical reactions. This can be phrased as a maximum common edge subgraph problem with the constraint that transition states must have cyclic structure. We describe a search tree method well suited for small edit distance and an integer linear program best suited for general instances and demonstrate that it is feasible to compute atom-atom maps at large scales using a manually curated database of biochemical reactions as an example. In this context we address the network completion problem.

1 Introduction

The individual records in databases of chemical reactions typically describe, apart from more or less detailed meta-information, the transformation of a set of educts into a set of products [30, 31]. Both the product and the educt molecules have representations as labeled graphs, where vertices designate atoms and edges refer to chemical bonds. Chemical reactions therefore may be understood as transformations of not necessarily connected graphs [5, 32]. Chemical graph transformations must respect the fundamental conservation principles of matter and charge and therefore imply the existence of a bijection between vertex sets (atoms) of the educts and products which is commonly known as the atom-atom map (AAM).

Chemical graph transformation are by no means arbitrary even when the conservation laws imposed by the underlying physics are respected. Instead, they conform to a large, but presumably finite, set of rules which in chemistry are collectively known as reaction mechanism and “named reactions”. A chemical reaction partitions the sets of atoms and bonds of the participating molecules into a *reaction center* comprising the bonds that change during the reactions and their incident atoms, and an remainder that is left unchanged. By virtue of being a bijection of the vertex (atom) sets, the AAM unambiguously determines the bonds that differ between educt and product molecules and thus it identifies the reaction center. The restriction of a chemical transformation to the reaction center, on the other hand, serves as minimal description of the underlying reaction rule.

The task to infer transformation rules from empirical chemical knowledge therefore would be greatly facilitated if each known reactions, i.e., each concrete pair of educt and product molecules would imply a unique graph transformation. Unfortunately, the true AAM is unknown in general, and even where the chemical mechanism, and thus the actual graph transformation, has been reported in the chemical literature, this information is in general not stored together with the educt/product pair in a database. The inference of chemical reaction mechanisms therefore requires that we first solve the problem of inferring AAMs for the known chemical reactions.

Several computational methods for the AAM problem have been devised and tested in the past [9]. The most common formulations are variants of the maximum common subgraph (isomorphism) problem [13]. In the NP-complete Maximum Common Edge Subgraph (MCES) variant an isomorphic subgraphs of both the educt and product graph with a maximal number of edges is identified. An alternative formulation as Maximum Common Induced Subgraph (MCIS) problem [1] is also NP complete. Algorithmic solutions decompose the molecules until only isomorphic sub-graphs remain [1, 11]. In the context of graph transformation systems, few methods to infer transformation rules have been published [20], and none applicable in the context of AAMs.

Neither solutions of MCES nor MCIS necessarily describe the true atom map, however. There is no reason why the re-organization of chemical bonds in a chemical reaction should maximize a subgraph problem. Instead, they follow strict rules that are codified, e.g., in the theory of imaginary transition states (ITS) [16, 18]. There is only a limited number of ITS “layouts” for single step reactions, corresponding to the cyclic electron redistribution pattern usually involving less than 10 atoms [19]. In a most basic case, an elementary reaction,

the broken and newly formed bonds form an alternating cycle of a length rarely exceeding 6 or at most 8 [18]. In [23] we made use of this chemical constraint to devise a Constraint Programming approach for elementary homovalent reactions, i.e., those chemical transformations that do not change the charge and oxidation state of an atom. Here, we use an extended representation of chemical graphs that explicitly represents lone pairs and bond orders; in this manner the graph representation incorporates more detailed chemical information.

Advances in bioinformatics technologies made it possible to infer large-scale metabolic networks automatically from genomic information [14, 6, 25]. These network models, however, suffer from structural gaps in pathways [7, 28], caused by orphan metabolic activities, for which no sequences are known and which cannot be inferred from genomic data. Thus there is an urgent need to infer missing metabolic reactions by other means. We illustrate the potential of AAM for the discovery of novel metabolic reactions. To this end we determine whether chemically plausible AAMs can be found connecting hypothetical educt/product pairs each consisting of one or two known metabolites.

2 Chemical Reactions are Cyclic

We model each molecule as a labeled, edge-weighted graph with loops. While the graph model used here is similar to most other formalizations of chemical graphs, it differs in several subtle, but important, details, such as the way charges and lone pairs are modeled:

Definition 1 (Molecule Graph). *A molecule graph $G = (V, E, l, w)$ is a labeled, edge-weighted, undirected graph with loops. The label function $l: V \cup E \rightarrow \Sigma_V \cup \Sigma_E$ denotes vertex and edge labels, and the weight function $w: E \rightarrow \mathbb{Z}$ denotes the weight of edges. These are assigned so that*

- *Atoms are vertices, with labels denoting which type of atom.*
- *Bonds are edges, with labels denoting the bond type and a weight encodes the number of involved electron pairs. Hence 1, 2, and 3 corresponds to single, double and triple bonds.*
- *Lone pairs, i.e., pairs of non-bonding electrons, are modeled by loops. Again the weight refers to the number of lone pairs.*
- *Charges are modeled using a single special vertex together with edges from this special vertex to the charged atoms. The edge weight equals the atom’s charge.*
- *Free radicals, single non-bonding electrons, are modeled using a single special vertex together with edges from this special vertex to the atom with the free radical. The edge weight equals the number of free radicals.*
- *Aromatic complexes are modeled by adding a special vertex for each aromatic complex in the molecules. Each atom participating in the aromatic complex has an edge to the special vertex with weight equal to the number of electrons at the atom taking part in the aromatic complex. The aromatic bonds themselves are edges with weight one, but are distinguished from single bonds by the edge label.*

See Fig. 1 for example of molecule graph. See Fig. 6 in Appendix D for an example of how the modeling of aromatic complexes works.

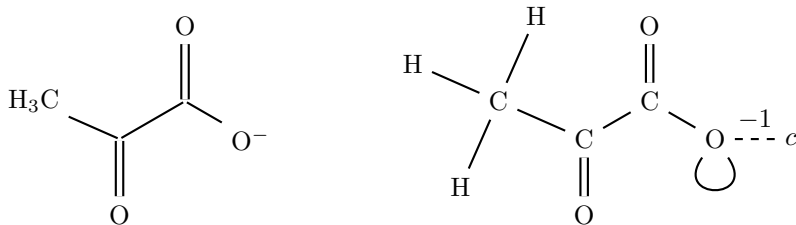


Figure 1: Usual depiction and molecule graph for pyruvate. Edge labels omitted. Edge weights shown by number of parallel edges (except where negative).

In the following two definitions it will be convenient to consider instead of E the set E^* of all possible edges on V with edges in $e \in E^* \setminus E$ having weight $w(e) = 0$.

Definition 2 (Atom-Atom Mapping). *Given two molecule graphs $G_1 = (V_1, E_1, l_1, w_1)$ and $G_2 = (V_2, E_2, l_2, w_2)$, an atom-atom mapping from G_1 to G_2 is a bijection $\psi: V_1 \rightarrow V_2$ that preserves vertex labels, i.e., $l_1(v) = l_2(\psi(v))$ for all $v \in V_1$. With ψ we associate the cost $c[\psi] = \sum_{e \in E_1^*} |w_2(\psi(e)) - w_1(e)|$.*

The cost measures the total number of electron pairs by which G_1 and G_2 differ w.r.t. to a given AAM. Minimizing $c(\psi)$ can be seen as an edit problem [21, 27, 17] and is equivalent to the NP -hard MCEs problem [2, 9, 13, 26, 4]. Here we are only interested in MCEs instances that correspond to balanced chemical reactions. The complexity results, however, also remains valid also in this case. Next we investigate in some more detail what exactly changes between G_1 and G_2 when an AAM ψ is fixed.

Definition 3 (Transition State). *The transition state of an AAM $\psi: G_1 \rightarrow G_2$ is the edge weighted graph $T_\psi = (V_\psi, E_\psi, w_\psi)$ where $E_\psi = \{e \in E_1^* \mid w_1(e) \neq w_2(\psi(e))\}$, $w_\psi(e) = w_2(\psi(e)) - w_1(e)$, and $V_\psi \subseteq V_1$ are all vertices incident to edges in E_ψ .*

By construction of molecule graphs, the weight of each edge is the number of valence electrons. The atom type, i.e., the label of a vertex determines the weighted degree $d_w(v) = \sum_{e \in \delta(v)} w(e)$. Here, loops are counted twice. This reflects that the two electrons per bond order are shared between the incident atoms, while both electrons of a lone pair belong to the same atom. As a consequence, $d_w(v)$ is invariant under all chemically acceptable atom maps. This restriction has important consequences for the structure of transition states:

Proposition 1 (Cyclic Transition States). *The transition state T_ψ of an AAM ψ can be decomposed into a collection of (not necessarily vertex disjoint) cycles C_1, C_2, \dots, C_k with weights $w_{C_1}, w_{C_2}, \dots, w_{C_k}$ that are alternating between $+1$ and -1 along the cycles such that $w_\psi(e) = \sum_{i=1}^k w_{C_i}(e)$ for all $e \in E_\psi$.*

Proof. Since AAMs preserve vertex labels and vertex labels imply weighted degree the “zero-flux condition” $\sum_{e \in \delta(v)} w_\psi(e) = 0$ holds for all $v \in V_\psi$. We

consider the following algorithm to construct a cycle C . Starting from a vertex v we choose an $\{v, u\}$, with $w_\psi(\{v, u\}) > 0$, decrement $w_\psi(\{v, u\})$ by one and add $\{v, u\}$ to C . The vertex u must be incident to an edge $\{u, w\}$ with $w_\psi(\{u, w\}) < 0$, since otherwise the weighted valence would not be constant under ψ . We increase $w_\psi(\{u, w\})$ by one and add $\{u, w\}$ to C . The process is repeated until we return to v , which is guaranteed by the finiteness of V . Clearly, C is an Eulerian graph, i.e., all its vertex degrees are even. The procedure is repeated until no edges with $w_\psi \neq 0$ is left. If C contains a vertex with degree larger than two, we repeat the procedure recursively on C until we are left with elementary cycles only. \square

The (weighted) degree $\delta_\psi(v) := \sum_{e:v \in e} |w_\psi(e)|$ of a vertex in T_ψ is even because in each step of the proof the value of $\delta_\psi(v)$ is reduced by 2. Thus T_ψ is a generalization of an Eulerian graph, and Prop. 1 is the corresponding variant of Veblen’s theorem [29], which states that a graph is Eulerian if and only if it is an edge-disjoint union of cycles.

3 Finding Atom-Atom Mappings

The cyclic nature of the transition states established in Prop. 1 inspires two methods for finding minimum cost AAMs described below. The idea was used in [23] in a much more restrictive chemical setting.

3.1 AltCyc — A Search Tree Approach

The idea of **AltCyc** is to construct a candidate transition state with a given cost ℓ in a stepwise fashion and to simultaneously map V_1 to V_2 . The search for transition states proceeds depth first. The validity of a candidate is then checked by testing whether $G_1 \setminus E_\psi$ and $G_2 \setminus \psi(E_\psi)$ are isomorphic. Finally, the parameter ℓ is increased until a valid mapping is found. A recursive definition of **AltCyc** is given as Algorithm 1.

Algorithm 1 **AltCyc**(ψ, P, k, σ)

```

if  $k = 1$  then
  if  $w_1(P.\text{head}, P.\text{tail}) + \sigma = w_2(\psi(P.\text{head}), \psi(P.\text{tail}))$  then
    Complete( $\psi, P$ )
  else
    for  $i \in V_1 \wedge i \notin \text{dom}(\psi)$  do
      for  $p \in V_2 \wedge p \notin \text{range}(\psi)$  do
        if  $l_1(i) = l_2(p) \wedge w_1(P.\text{head}, i) + \sigma = w_2(\psi(P.\text{head}), p)$  then
           $\psi \leftarrow \psi \cup \{i \mapsto p\}$ 
          AltCyc( $\psi, P.\text{append}(i), k - 1, -1 \cdot \sigma$ )

```

To explain the algorithm, we first restrict ourself to mappings with transition states consisting of a single elementary cycle. The four parameters are a partial atom-atom mapping candidate ψ , the partial transition state P constructed so far encoded as a list of vertices from V_1 , and the number k of edges still to be identified, and the variable $\sigma \in \{-1, 1\}$ that determines whether the current step will add or remove weight.

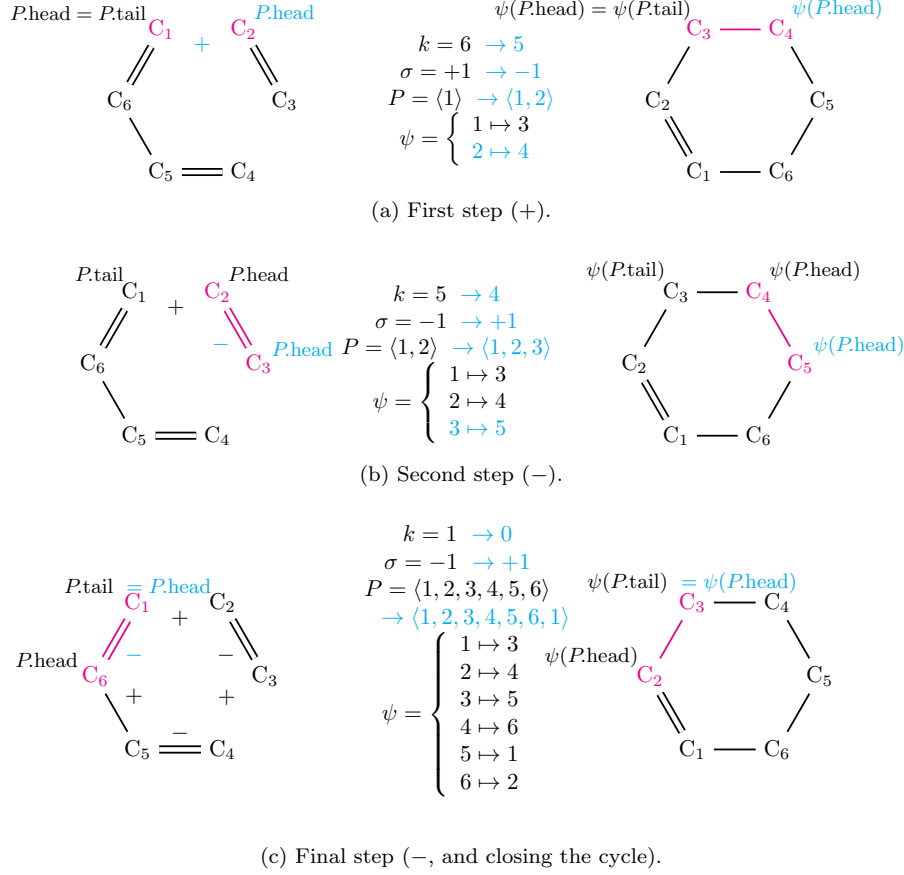


Figure 2: Stepwise execution of **AltCyc**. Cyan marks the changes within the step. Magenta marks the considered edges and incident vertices.

The search in **AltCyc** starts from all pairs (i, p) with $i \in V_1$ and $p \in V_2$ with $l_2(p) = l_1(i)$; the map ψ is initialized $\psi(i) = p$ and the path starts with $P = \{i\}$. W.l.o.g., the first step is a positive change of weight, i.e., $\sigma = 1$. In each step in the algorithm, a new pair $(i, p) \in V_1 \times V_2$ with matching labels is found and if the $w_1(\{P.\text{head}, i\})$ and $w_2(\{\psi(P.\text{head}), p\})$ differ by exactly one, i is appended to P , ψ is extended such that $\psi(i) = p$ and the algorithm is called again with k replaced by $k - 1$. If $k = 1$ has been reached, it only remains to close the alternating cycle. If this is possible, the candidate transition state is extended to a full AAM where no further changes are allowed. To this end, a graph isomorphism algorithm is used. We use **VF2** [10] in procedure **Complete** (see Appendix C) because it has the added benefit of using data structures that are similar to those used in other parts of **AltCyc**. The first two and the last step of **AltCyc** applied to a Diels-Alder reaction are shown in Fig. 2.

In order to handle transition states that are connected but not elementary cycles, as the case of a bi-cyclic or coarctate reaction [18], we modify **AltCyc** to allow weight differences larger than one. Such vertices must then be revisited. In addition, we disallow using the same edge with different signs of σ because

a pair of such steps would cancel. The modified approach is outlined in Algorithm 2. The key point is that we now need to keep track of the weight changes, $w_P(e)$, that we have already made along an edge e (found using the procedure **WeightAlongPath**, see Appendix C). The condition for acceptable weight differences becomes $w_1(e) + w_P(e) + \sigma \leq w_2(\psi(e))$ if a bond is added ($\sigma = 1$), and $w_1(e) + w_P(e) + \sigma \geq w_2(\psi(e))$ for bond subtraction ($\sigma = -1$).

Algorithm 2 $\text{AltCyc}^*(\psi, P, k, \sigma)$

```

// As AltCyc...
for  $(i, p) \in V_1 \times V_2$  with  $l_1(i) = l_2(p)$  do
  if  $i \notin \text{dom}(\psi) \wedge p \notin \text{range}(\psi)$  then
    // As AltCyc, but using  $\leq$  and  $\geq$  ...
  else if  $\psi(i) = p$  then
     $w_P \leftarrow \text{WeightAlongPath}(\{P.\text{head}, i\}, P)$ 
    if  $w_P \geq 0 \wedge \sigma = 1$  then
      if  $w_1(P.\text{head}, i) + w_P + \sigma \leq w_2(\psi(P.\text{head}), p)$  then
         $\text{AltCyc}^*(\psi, P.\text{append}(i), k - 1, -1 \cdot \sigma)$ 
    else if  $w_P \leq 0 \wedge \sigma = -1$  then
      // Symmetric case...

```

There is no guarantee that the transition state is connected. To accommodate disconnected transition states it suffices to replace the path P by a list of paths, where the last path is the current path and all previous paths are kept in order to correctly calculate $w_P(e)$. If a path closes before $k = 0$ is reached, the current cycle is completed and the algorithm restarts to build new path from another initial vertex.

The stepwise approach in **AltCyc** naturally allows for an elucidation of the mechanism underlying an AAM found by the algorithm. In Fig. 3 the automatic inference of such a mechanism is illustrated. Each step in the figure, the usual way of drawing arrow pushing diagrams, corresponds to two steps in **AltCyc**.

Taken together, **AltCyc** uses $O((n^2)^k) = O(n^{2k})$ recursive calls, where $n = |V_1| = |V_2|$. Exploiting the fact that only edges to the special vertex for a charge can be negative, this reduces to $O(n^{k+l})$, where l is the number of components in the transition state candidate, because it suffices to examine only the $O(1)$ edges incident to $P.\text{head}$ or $\psi(P.\text{head})$ depending on whether we are making a negative or positive step in the algorithm. In addition, **AltCyc** incurs the cost of the graph isomorphism check for completing the mapping.

In practice, however, the runtime is much lower since vertex labels must match. The runtime nevertheless still depends heavily on k , and thus the length of the optimal mapping of the instance. However, as discussed k can be assumed to be small for the case of inferring chemical transformation rules. Due to depth first strategy, the memory consumption of **AltCyc** is $O(n)$.

3.2 ILP2 — An Integer Linear Program

The AAM problem can also be phrased as an ILP. We use binary variables m_{ip} to encode the mapping ψ as $m_{ip} = 1$ iff $\psi(i) = p$ and $m_{ip} = 0$ for all other combinations of i and p . To enforce that ψ is vertex label preserving we

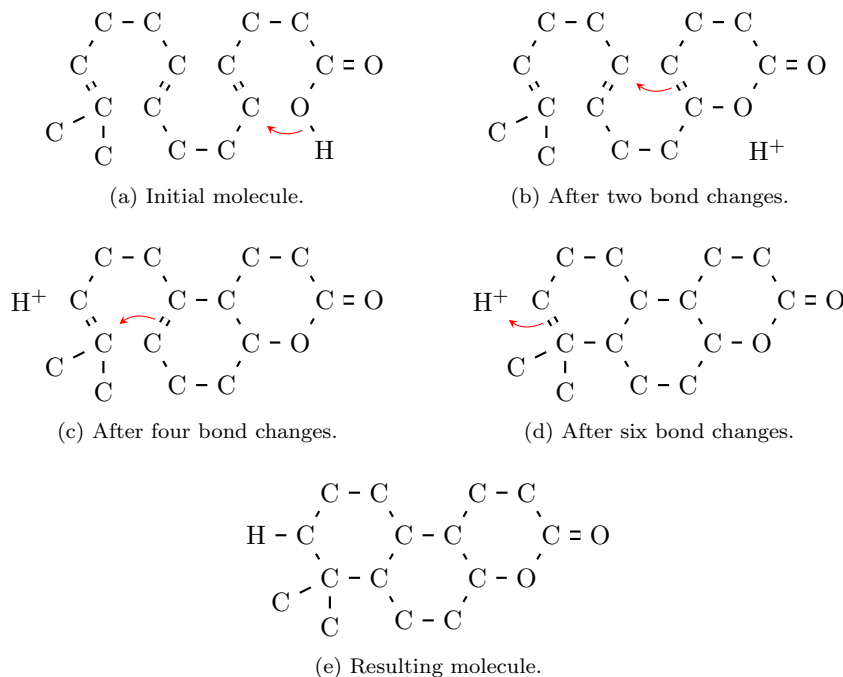


Figure 3: An example AAM for Stork’s cyclisation of farnesyl acetic acid to ambreinolide [33]. Note that only the single hydrogen in the transition state is shown, and while it is assumed in the model that it is the same hydrogen leaving and later entering, in actual chemistry it is a different hydrogen.

set $m_{ip} = 0$ for $l_1(i) \neq l_2(p)$, and to ensure ψ is a bijection we formulate the following linear constraints.

$$\forall i \in V_1: \sum_{p \in V_2} m_{ip} = 1 \quad \text{and} \quad \forall p \in V_2: \sum_{i \in V_1} m_{ip} = 1$$

The most obvious way to proceed would be to keep track of the mapping between the edge sets using either binary variables describing whether a bond is mapped or not as in [15], or integer variables that denote the weight change if a bond is mapped, and zero for unmapped bonds. For such variables we would need $O(|V|^4)$ constraints, however. Empirically we found that ILP-solvers quickly run out of memory and become very slow for such a model.

Though there already exist ILP formulations of similar problems with only $O(|V|^2)$ constraints [22], obtained by exploiting the sparseness of molecule graphs, we propose a new ILP formulation based on the Kaufmann and Broeckx linearization of the quadratic assignment problem [8], which also needs only $O(|V|^2)$ constraints.

We introduce integer variables $c_{ip}^+ \in \mathbb{N}_0$ and $c_{ip}^- \in \mathbb{N}_0$ that model the positive and negative weight changes respectively of all edges incident to vertex $i \in V_1$ if $\psi(i) = p$. Both c_{ip}^+ and c_{ip}^- are zero for all other combinations of i and p . Making use of the fact that weight changes are balanced, i.e. $\sum_{e \in \delta(v)} w_\psi(e) = 0$

for all $v \in V_\psi$, we can use the following constraint for all $i \in V_1$:

$$\sum_{p \in V_2} c_{ip}^+ = \sum_{p \in V_2} c_{ip}^-$$

We also substitute them in the objective function:

$$obj = \sum_{(i,p) \in V_1 \times V_2} c_{ip}^+ + \sum_{(i,p) \in V_1 \times V_2} c_{ip}^-$$

Since the change variables are included in the objective function they will implicitly be constrained from above. In order to constrain them from below we use the following constraints for all $(i, p) \in V_1 \times V_2$:

$$c_{ip}^+ \geq (m_{ip} - 1) \cdot M + \sum_{(j,q) \in V_1 \times V_2} m_{jq} \cdot \max\{0, w_2(\{p, q\}) - w_1(\{i, j\})\}$$

$$c_{ip}^- \geq (m_{ip} - 1) \cdot M + \sum_{(j,q) \in V_1 \times V_2} m_{jq} \cdot \max\{0, w_1(\{i, j\}) - w_2(\{p, q\})\}$$

where M is a suitably large constant. It suffices to set M to the largest weighted degree to void the constraint when $m_{ip} = 0$. The first term voids the constraints if $m_{ip} \neq 1$. The sums correspond to the sum of all positive (negative) changes of edges incident to i and p respectively, if indeed these edges are mapped to each other.

Unlike **AltCyc** we have little control over intermediate steps in the reaction, but using **ILP2** we have much freedom to modify the cost model used. Assuming we have an integer linear programming solver available **ILP2** takes very little time to implement.

3.3 Enumeration of All Optimal Atom-Atom Mappings

So far we have focused on the problem of finding a single AAM. The solution of the optimization problem is in general not unique, however. A particular problem in this context are symmetries of the educt or product molecules, because this may bloat the number of AAMs. We are therefore interested only in nonequivalent AAMs.

Definition 4 (Equivalent Atom-Atom Mappings). *For a given AAM define $G_\psi = (V_\psi, E_\psi, l_\psi)$ with vertex set $V_\psi = V_1$, edge set $E_\psi = E_1 \cup \psi^{-1}(E_2)$, and label function $l_\psi(x) = (l_1(x), l_2(\psi(x)))$. If $x \notin \text{dom}(l_i)$ then $l_i(x) = \varepsilon_i$, where ε_i is some label not in $\text{range}(l_i)$, denoting a non-edge. We say two atom-atom mappings, ψ and φ are equivalent if the graphs G_ψ and G_φ are isomorphic.*

Now, let us consider whether a transition state candidate of an atom-atom mapping uniquely defines the full mapping.

Definition 5 (Completion of Partial Mapping). *Given a partial AAM $\psi': A \subset V_1 \rightarrow B \subset V_2$, a completion of ψ' is an AAM such that $\psi|_A = \psi'$ and outside A , ψ preserves all properties of G_1 and G_2 .*

Note that such a completion need not exist for a given partial AAM.

Proposition 2 (Partial Mapping). *If ψ and φ are two completions of a partial AAM ψ' , ψ and φ are equivalent.*

Proof. Consider the two AAMs ψ and φ and their associated graphs G_ψ and G_φ . By assumption, they are both completions of the same partial AAM ψ' ; therefore the two induced sub-graphs $G_\psi[\text{dom}(\psi')]$ and $G_\varphi[\text{dom}(\psi')]$ are identical. Consider the subgraphs $G' := G_\psi \setminus E(G_\psi[\text{dom}(\psi')])$ and $G'' := G_\varphi \setminus E(G_\varphi[\text{dom}(\psi')])$ without edges in $G_\psi[\text{dom}(\psi')]$. G' and G'' both are identical to $G_1 \setminus E(G_1[\text{dom}(\psi')])$ if only considering the labels from l_1 in each of G_ψ and G_φ . As both ψ and φ preserve all properties of G_1 and G_2 outside $\text{dom}(\psi')$, the labels from l_2 are always identical to the labels from l_1 outside $\text{dom}(\psi')$.

Thus G_ψ and G_φ are isomorphic and by definition ψ and φ are equivalent. \square

Prop. 2 can be applied in different ways. In **AltCyc** it shows we only have to complete each candidate transition state once in order to enumerate all mappings. In **ILP2** it can be used to exclude solutions based on mapping variables defining the transition states instead of all mapping variables.

4 Results

The RHEA [24] database (v. 50), which provides access to a large set of expert-curated biochemical reactions, has been used to test our suggested AAM algorithms, and to underline the necessity of graph transformation methods for network completion. We exclude all reactions with unspecified repeating units and wildcards, resulting in a set of 19753 reactions involving a set, M , of 3786 non-isomorphic molecular graphs. We performed a statistical analysis of RHEA, that shows how often molecules are used in the reaction listed in the database, and how many non-isomorphic isomers are stored in RHEA. Interestingly, terpene chemistry [12] clearly dominates the high frequency isomers (see Appendix A). Due to space limitations, we focus on a brief runtime analysis and network completion results. As **AltCyc** constructs solutions in a stepwise fashion, a chemical mechanism explaining the bond changes as subsequent transformations is naturally inferred. An example for a mechanistic inference of Stork’s cyclisation of farnesyl acetic acid to ambreinolide [33] is given in Fig. 3.

Runtime. We compared **AltCyc**, **ILP2**, and a naïve ILP-implementation with $O(n^4)$ constraints, **ILP4**, with regard to their ability of enumerating all non-equivalent AAMs within a fixed runtime (see appendix B). We found that **ILP2** drastically outperforms the naïve ILP-implementation and also is systematically more efficient than **AltCyc**. The latter has a (small) advantage for instances with small transition states. For both methods we see an exponential decline in ratio of quickly solved instances as size of instances grow, this corresponds well with the expected exponential runtime.

Network Completion. Databases of metabolic networks are by no means complete because the enzymes catalyzing many of the reactions in particular in the so-called secondary metabolism have remained unknown. Furthermore, for almost one third of the known metabolic activities, no protein sequences are known that could encode the corresponding enzyme. *Network completion* is an important task to fill gaps i.e. missing reaction steps, in genome-scale metabolic networks. Reaction perception, i.e. finding AAMs, is the only technique capable

of finding possible candidates for the missing reactions, where homology based methods fail, due to lack of data.

Inferring all candidate *2-to-2* reactions addresses this issue by determining for all disjoint pairs A, B of multisets A (one or two educt molecules, potentially isomorphic) and multisets B (one or two product molecules, also potentially isomorphic), whether there is a chemically plausible reaction transforming A to B . By Prop. 1, *any* reaction satisfying mass and charge balance has a cyclic transition state.

Let $R_{2,2}$ denote the set of all sets $\{A, B\}$ such that A and B are disjoint multi-subsets of the set of molecules M , both of size at most 2 with A and B containing the same vertex labels, charges, etc. The set of test instances $R_{2,2}$ of *2-to-2* reactions that satisfy mass balance can be extracted from a database with molecule set M in time $O(|M|^2 \log |M| + |R_{2,2}|)$ using Algorithm 5 (see Appendix C). We obtain a set of $|R_{2,2}| = 114,429,849$ balanced reaction candidates with at most two molecules on either side of the reaction.

It is not feasible to test 100 million candidates for chemical feasibility in an exact manner. Using the length of the transition state as a filter, however, will remove implausible candidates as well as multi-step mechanisms. The length of the transition state can be bounded in both **AltCyc** as well as **ILP2**. We used **AltCyc** because of its performance advantage with small transition states. In a random sample of 10,000 instances drawn from $R_{2,2}$ we found 34, 59, and 167 reactions with transition states of length 4, 6, and 8, respectively. Extrapolating from this sample we have to expect approximately three million candidate reactions with AAMs that will need to be examined in more detail. Clearly this number is too large for a biochemical network. Further pruning of the candidate list will thus require additional information, e.g., on the energetics of the reactions and on these reaction mechanisms plausibly catalyzed by enzymes. However, it underlines the need for graph transformation techniques for computing realistic candidate sets.

5 Conclusion

Graph transformation systems have great potential as a model of chemistry in particular in the context of large reaction networks. Their practical usefulness, however, stands and falls with the ability to produce collections of transformation rules that closely reflect chemical reality. We have shown here that the extraction of AAMs from educt/product pairs is a necessary first step because the restriction of the graph transformation to the reaction center, which is uniquely determined by the AAM, provides a minimal description of the corresponding reaction rule. We have shown formally that it is not sufficient to solve a general graph editing problem. Instead, the cyclic nature of the transition states must be taken into account as additional constraints. With **AltCyc** and **ILP2** we have introduced two complementary approaches to solve this chemically constrained maximum subgraph problem. The constructive **AltCyc** approach performs better on short cycle instances. If more complex transition states need to be considered or if flexibility in the cost function is required **ILP2** becomes the method of choice.

Advances in high-throughput sequencing technologies drives the reconstruction of organism-specific large-scale metabolic networks from genomic sequence

information. Reaction perception, as illustrated here on the Rhea database, is currently the only computational technique to suggest missing reactions in the reconstructed networks once the methods of comparative genomics to infer enzyme activities are exhausted. We have demonstrated here that efficient computation of AAMs serves as first effective step. Much remains to be done, however. Most importantly, the AAM determines only a minimal reaction rule confined to the reaction center. The feasibility of chemical reactions, however, also depends on additional context in the vicinity of the reaction center. While graph grammar systems readily accommodate non-trivial context [5, 3], we have yet to develop methods to infer the necessary contexts from the huge body of chemical reaction knowledge. Once this is solved, such more elaborate rules will form a highly efficient filter for the candidate AAMs. In this context the stepwise construction of the transition state in `AltCyc` holds further promise: context information could be used efficiently already in the AAM construction step to prune its search tree, simultaneously increase the chemical realism of the solutions and its computational efficiency.

Acknowledgments

This work was supported in part by the Volkswagen Stiftung proj. no. I/82719, and the COST-Action CM1304 “Systems Chemistry” and by the Danish Council for Independent Research, Natural Sciences.

References

- [1] T Akutsu. Efficient extraction of mapping rules of atoms from enzymatic reaction data. *J. Comp. Biol.*, 11:449–62, 2004.
- [2] Tatsuya Akutsu and Takeyuki Tamura. A polynomial-time algorithm for computing the maximum common connected edge subgraph of outerplanar graphs of bounded degree. *Algorithms*, 6(1):119, 2013.
- [3] Jakob Lykke Andersen, Christoph Flamm, Daniel Merkle, and Peter F. Stadler. 50 shades of rule composition. In Franois Fages and Carla Piazza, editors, *Formal Methods in Macro-Biology*, volume 8738 of *Lecture Notes in Computer Science*, pages 117–135. Springer International Publishing, 2014.
- [4] Laura Bahiense, Gordana Mani, Breno Piva, and Cid C. de Souza. The maximum common edge subgraph problem: A polyhedral investigation. *Discrete Applied Mathematics*, 160(18):2523 – 2541, 2012. V Latin American Algorithms, Graphs, and Optimization Symposium Gramado, Brazil, 2009.
- [5] Gil Benkő, Christoph Flamm, and Peter F. Stadler. A graph-based toy model of chemistry. *J. Chem. Inf. Comput. Sci.*, 43:1085–1093, 2003. presented at *MCC 2002*, Dubrovnik CRO, June 2002; SFI # 02-09-045.
- [6] Matthew B. Biggs and Jason A. Papin. Metabolic network-guided binning of metagenomic sequence fragments. *Bioinformatics*, 2015.
- [7] Rainer Breitling, Dennis Vitkup, and Michael P Barrett. New surveyor tools for charting microbial metabolic maps. *Nature Rev Microbiol*, 6:156–161, 2008.
- [8] Rainer E. Burkard, Eranda ela, Panos M. Pardalos, and Leonidas S. Pitsoulis. The quadratic assignment problem. In Ding-Zhu Du and Panos M. Pardalos, editors, *Handbook of Combinatorial Optimization*, pages 1713–1809. Springer US, 1999.
- [9] W. L. Chen, D. Z. Chen, and K. T. Taylor. Automatic reaction mapping and reaction center detection. *WIREs Comput Mol Sci*, 3:560–593, 2013.
- [10] Luigi P. Cordella, Foggia Pasquale, Carlo Sansone, and Mario Vento. A (sub)graph isomorphism algorithm for matching large graphs. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(10):1367–1372, 2004.
- [11] J.D. Crabtree, D.P. Mehta, and T.M. Kouri. An open-source java platform for automated reaction mapping. *J Chem Inf Model*, 50:1751–1756, 2010.

- [12] Jörg Degenhardt, Tobias G Köllner, and Jonathan Gershenzon. Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. *Phytochem*, 70:1621–1637, 2009.
- [13] H.-C. Ehrlich and M. Rarey. Maximum common subgraph isomorphism algorithms and their applications in molecular science: a review. *WIREs Comput Mol Sci*, 2011. doi:10.1002/wcms.5.
- [14] Adam M Feist, Markus J Herrgård, Ines Thiele, Jennie L Reed, and Bernhard Ø Palsson. Reconstruction of biochemical networks in microorganisms. *Nature Rev Microbiol*, 7:129–143, 2009.
- [15] E. L. First, C. E. Gounaris, and C. A. Floudas. Stereochemically consistent reaction mapping and identification of multiple reaction mechanisms through integer linear optimization. *J Chem Inf Model*, 52:84–92, 2012.
- [16] S. Fujita. Description of organic reactions based on imaginary transition structures. 1. introduction of new concepts. *J. Chem. Inf. Comput. Sci.*, 26:205–212, 1986.
- [17] Xinbo Gao, Bing Xiao, Dacheng Tao, and Xuelong Li. A survey of graph edit distance. *Pattern Analysis and Applications*, 13(1):113–129, 2010.
- [18] J. B. Hendrickson. Comprehensive system for classification and nomenclature of organic reactions. *J Chem Inf Comput Sci*, 37:852–860, 1997.
- [19] Rainer Herges. Organizing principle of complex reactions and theory of coarctate transition states. *Angewandte Chemie Int Ed*, 33:255–276, 1994.
- [20] Eric Jeltsch and Hans-Jörg Kreowski. *Graph Grammars and Their Application to Computer Science: 4th International Workshop Bremen, Germany, March 5–9, 1990 Proceedings*, chapter Grammatical inference based on hyperedge replacement, pages 461–474. Springer Berlin Heidelberg, Berlin, Heidelberg, 1991.
- [21] D. Justice and A. Hero. A binary linear programming formulation of the graph edit distance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(8):1200–1214, Aug 2006.
- [22] Mario Latendresse, Jeremiah P. Malerich, Mike Travers, and Peter D. Karp. Accurate atom-mapping computation for biochemical reactions. *J Chem Inf Model*, 52:2970–2982, 2012.
- [23] Martin Mann, Feras Nahar, Norah Schnorr, Rolf Backofen, Peter F. Stadler, and Christoph Flamm. Atom mapping with constraint programming. *Alg. Mol. Biol.*, 9:23, 2014.
- [24] Anne Morgat, Kristian B. Axelsen, Thierry Lombardot, Rafael Alcantara, Lucila Aimó, Mohamed Zerara, Anne Niknejad, Eugeni Belda, Nevila Hyka-Nouspikel, Elisabeth Coudert, Nicole Redaschi, Lydie Bougueleret, Christoph Steinbeck, Ioannis Xenarios, and Alan Bridge. Updates in reea a manually curated resource of biochemical reactions. *Nucleic Acids Research*, 43(D1):459–464, 2015.

- [25] Sylvain Prigent, Guillaume Collet, Simon M. Dittami, Ludovic Delage, Floriane Ethis de Corny, Olivier Dameron, Damien Eveillard, Sven Thiele, Jeanne Cambefort, Catherine Boyen, Anne Siegel, and Thierry Tonon. The genome-scale metabolic network of *ectocarpus siliculosus* (ectogem): a resource to study brown algal physiology and beyond. *The Plant Journal*, 80(2):367–381, 2014.
- [26] John W. Raymond and Peter Willett. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *Journal of Computer-Aided Molecular Design*, 16(7):521–533, 2002.
- [27] Kaspar Riesen and Horst Bunke. Approximate graph edit distance computation by means of bipartite graph matching. *Image and Vision Computing*, 27(7):950 – 959, 2009. 7th IAPR-TC15 Workshop on Graph-based Representations (GbR 2007).
- [28] Torsten Schaub and Sven Thiele. *Logic Programming: 25th International Conference, ICLP 2009, Pasadena, CA, USA, July 14-17, 2009. Proceedings*, chapter Metabolic Network Expansion with Answer Set Programming, pages 312–326. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.
- [29] O Veblen. An application of modular equations in analysis situs. *Ann. Math.*, 14:86–94, 1912.
- [30] Wendy A. Warr. A short review of chemical reaction database systems, computer-aided synthesis design, reaction prediction and synthetic feasibility. *Molecular Informatics*, 33:469–476, 2014.
- [31] U Wittig, M Rey, R Kania, M Bittkowski, L Shi, M Golebiewski, A Weidemann, W Müller, and I Rojas. Challenges for an enzymatic reaction kinetics database. *FEBS J.*, 281:572–582, 2014.
- [32] Maneesh K. Yadav, Brian P. Kelley, and Steven M. Silverman. *Graph Transformations: Second International Conference, ICGT 2004, Rome, Italy, September 28–October 1, 2004. Proceedings*, chapter The Potential of a Chemical Graph Transformation System, pages 83–95. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.
- [33] Ryan A Yoder and Jeffrey N Johnston. A case study in biomimetic total synthesis: Polyolefin carbocyclizations to terpenes and steroids. *Chem Rev*, 105:4730–4756, 2005.

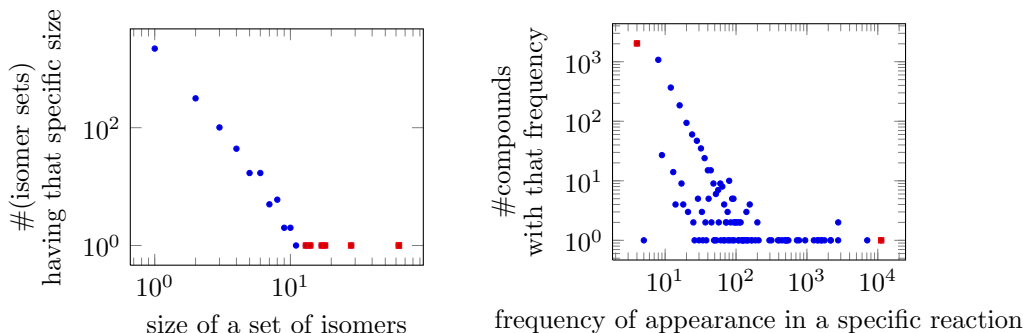


Figure 4: Distribution of isomers and frequency of participation in reactions in Rhea. Left plot shows a few sets of isomers are very large, while most compounds in Rhea are unique up to sum formula of those compounds. Right plot shows the frequency with which a compound participates in reactions.

A Statistical Analysis of Rhea

Of the $M = 3786$ non-isomorphic molecular graphs in RHEA, 2204 are identified uniquely by their sum formula. While 2030 of the molecules appear only in a minimum of 4 reactions, some compounds take part in a very large fraction of all reactions in RHEA, e.g., H^+ participates in 11,147 reactions, some of which are different descriptions of similar reactions where only the direction of the reaction differs, 5055 of these are truly distinct, adenosine di-, and tri-phosphate (and its derivatives), water, and dioxide each participate in more than 2000 reactions (depicted as red dots in Fig. 4 (right)). The maximum number of isomers (i.e., compounds that have the same sum formula but a non-isomorphic graph representation) is 63. The corresponding sum formula is $C_{15}H_{24}$. Interestingly, most of the large sets of isomers in RHEA are terpenes, condensates of identical five carbon atom building blocks. The terpenes form a combinatorial class of polycyclic ring-systems via complex sequences of cyclisation and isomerization reactions. Fig. 4 (left) summarizes the results (terpenes marked with red).

B Analysis of Runtime

As we are mainly interested in single step reactions, we restricted our algorithms to only look for connected, vertex-disjoint transition states during the comparison. Fig. 5 shows the fraction of instances where `AltCyc`, `ILP2` and a naïve ILP-implementation with $O(n^4)$ constraints, `ILP4`, are able to enumerate all non-equivalent atom-atom mappings for different instance size categories as well as absolute number of instances solved divided by solution size.

Only very few instances that are not completely solved within the first 60 seconds are solved within reasonable time (one hour). So there seems to be a sharp divide between easy and hard instances. From the plot in Fig. 5 (left) of the fraction of instances solved fast we observe an exponential decline in ratio of solved instances. This corresponds well with the expected exponential runtime of the algorithms.

As we restricted the solution set certain instances are proven infeasible by

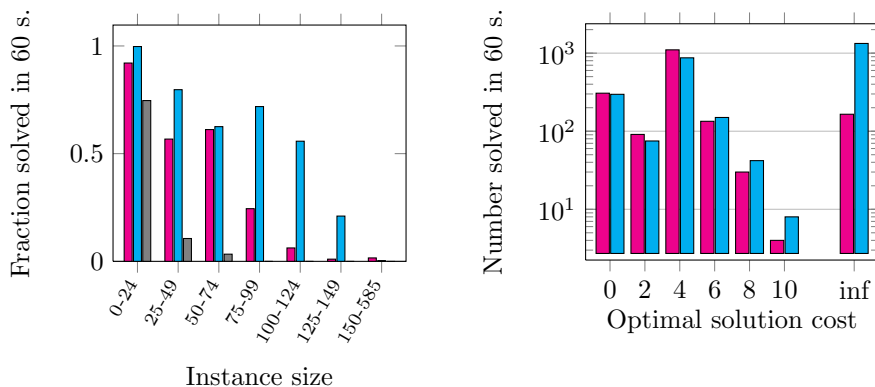


Figure 5: Fraction and number of instances where all optimal atom-atom maps are found in 60 seconds (user time) by instance size and optimal solution cost for AltCyc (magenta), ILP2 (cyan) and ILP4 (gray).

ILP2, while AltCyc will continue searching for solutions until the parameter k , the number of weight changes, gets arbitrarily high. We chose to deem instances where AltCyc found no solutions for $k \leq 10$ infeasible and terminate the search. These two classes of solutions are marked in the rightmost column in Fig. 5. Note that the performance of AltCyc on the infeasible class of instances depends heavily on the somewhat arbitrary choice of maximum k .

Both ILP models are implemented using CPLEX, an efficient state of the art MIP-solver. AltCyc and ILP2 has been tested on a total of 4295 Rhea instances, while ILP4 has only been tested on a subset of these of size 250.

C Algorithmic Details

For completeness we include pseudo-code for the sub-procedures used in the paper.

Pseudo-code for WeightAlongPath: In AltCyc* (see Algorithm 2) we need to find all previous changes to an edge $\{i, j\}$ currently under examination, $w_P(\{i, j\})$.

In Algorithm 3 we show how to do this in time $O(|P|)$, where $|P| \in O(k)$. It is possible to find $w_P(e)$ in constant time, but this would require much more complicated data structures or making changes to the graphs we work on and as k is in practice very small, this method is preferred.

To find $w_P(e)$ for a list of paths, add $w_P(e)$ for all paths in the list.

Algorithm 3 WeightAlongPath($\{i, j\}, P$)

```
 $w_P \leftarrow 0$ 
 $\sigma \leftarrow 1$ 
for  $i'$  from 0 to  $|P| - 2$  do
     $j' \leftarrow i' + 1$ 
    if  $\{i', j'\} = \{i, j\}$  then
         $w_P \leftarrow w_P + \sigma$ 
     $\sigma \leftarrow -1 \cdot \sigma$ 
```

Pseudo-code for Complete: When a transition state candidate ψ' is found we need to ensure it can be extended into a complete atom-atom mapping. This can be done as described in Algorithm 4. Note that the two graphs G_1 and G_2 are assumed implicitly known. The algorithm works both for a single path, P , or where P represents a list of paths.

The only non-trivial detail in Algorithm 4 is that it is not correct to remove all edges in the induced subgraph on the domain of ψ' , the weight change needs to be sufficient, and there may be unchanged cords to consider.

Algorithm 4 Complete(ψ', P)

```
for  $e \in P$  do {Here  $P$  is considered a set of edges}
     $w_P \leftarrow \text{WeightAlongPath}(e, P)$ 
    if  $w_P = w_2(\psi(e)) - w_1(e)$  then
        Remove  $e$  from  $G_1$  and  $\psi(e)$  from  $G_2$ 
    else
        fail
for  $(i, p) \in V_1 \times V_2$  where  $\psi'(i) = p$  do
    Relabel  $i$  and  $p$  to have identical, otherwise unique labels
return an isomorphism from  $G_1$  to  $G_2$ 
```

Finding 2-to-2 Candidates in $O(n^2 \log n)$ Comparisons. In order to generate all $O(n^4)$ candidate reactions with no more than two molecules in the educts or products we use Algorithm 5. A set of molecules, M , is given, as well as a method to obtain the distribution of atoms and charges of the molecules h , in practice some implementation of sparse vectors. We assume we keep pointers to the original molecules that resulted in each distribution, and we get these with the function *mol*.

Algorithm 5 2to2(M)

```
 $\mathcal{H} \leftarrow h(M) \cup \{\vec{0}\}$ 
generate  $H = \{h_1 + h_2 \mid (h_1, h_2) \in \mathcal{H} \times \mathcal{H} \wedge h_1 \leq h_2\}$  as an array
Sort( $H$ )
for  $i \leftarrow 1$  to  $\text{len}(H) - 1$  do
     $j \leftarrow i + 1$ 
    while  $j \leq \text{len}(H) \wedge H[i] = H[j]$  do
        output ( $\text{mol}(H[i]), \text{mol}(H[j])$ )
     $j \leftarrow j + 1$ 
```

The algorithm is dominated by one of two things, either the sorting of the length n^2 array H (where $n = |M|$), or the time to output candidates $k \in O(n^4)$, the resulting runtime is then $O(n^2 \log n + k)$.

D Example of Aromatic Structure

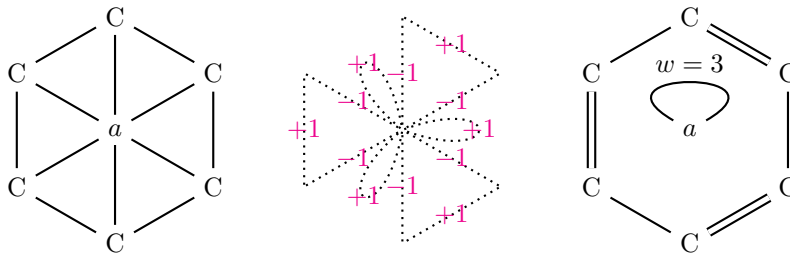


Figure 6: Illustration of modeling of aromatic cycle. Left is an aromatic cycle, right is the same cycle in Kukulé form. Edge labels are not shown, and edge weight is implied with multiple parallel lines. The figure in the middle depicts the alternating transition state between the two assuming AAM by position of atoms.

It is non-trivial that the model presented here of aromatic complexes will allow for AAMs with cyclic transition states, Fig. 6 illustrates how this can be done. We add special aromatic vertices with loops to either G_1 or G_2 to ensure the AAM is still a bijection and that a mapping is feasible.